


# Refined image quality assessment for color fundus photography based on deep learning

DIGITAL HEALTH  
Volume 10: 1–13  
© The Author(s) 2024  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/20552076231207582  
journals.sagepub.com/home/dhj



Tianjiao Guo<sup>1,2,3</sup> , Kun Liu<sup>4,5,6</sup>, Haidong Zou<sup>4,5,6</sup>, Xun Xu<sup>4,5,6</sup>,  
Jie Yang<sup>2</sup> and Qi Yu<sup>4,5,6</sup>

## Abstract

**Purpose:** Color fundus photography is widely used in clinical and screening settings for eye diseases. Poor image quality greatly affects the reliability of further evaluation and diagnosis. In this study, we developed an automated assessment module for color fundus photography image quality assessment using deep learning.

**Methods:** A total of 55,931 color fundus photography images from multiple centers in Shanghai and the public database were collected and annotated as training, validation, and testing data sets. The pre-diagnosis image quality assessment module based on the multi-task deep neural network was designed. The detailed criterion of color fundus photography image quality including three subcategories with three levels of grading was applied to improve precision and objectivity. The auxiliary tasks such as the localization of the optic nerve head and macula, the classification of laterality, and the field of view were also included to assist the quality assessment. Finally, we validated our module internally and externally by evaluating the area under the receiver operating characteristic curve, sensitivity, specificity, accuracy, and quadratic weighted Kappa.

**Results:** The “Location” subcategory achieved area under the receiver operating characteristic curves of 0.991, 0.920, and 0.946 for the three grades, respectively. The “Clarity” subcategory achieved area under the receiver operating characteristic curves of 0.980, 0.917, and 0.954 for the three grades, respectively. The “Artifact” subcategory achieved area under the receiver operating characteristic curves of 0.976, 0.952, and 0.986 for the three grades, respectively. The accuracy and Kappa of overall quality reach 88.15% and 89.70%, respectively, on the internal set. These two indicators on the external set were 86.63% and 88.55%, respectively, which were very close to that of the internal set.

**Conclusions:** This work showed that our deep module was able to evaluate the color fundus photography image quality using more detailed three subcategories with three grade criteria. The promising results on both internal and external validation indicated the strength and generalizability of our module.

## Keywords

Color fundus photography, deep learning, image quality assessment, screening, pre-diagnosis

Submission date: 2 February 2023; Acceptance date: 26 September 2023

<sup>1</sup>Institute of Medical Robotics, Shanghai Jiao Tong University, China

<sup>2</sup>Department of Automation, Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, China

<sup>3</sup>School of Biomedical Engineering, Shanghai Jiao Tong University, China

<sup>4</sup>Department of Ophthalmology, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, China

<sup>5</sup>National Clinical Research Center for Eye Diseases, Shanghai, China

<sup>6</sup>Shanghai Clinical Research Center for Eye Diseases, China

## Corresponding author:

Jie Yang, Department of Automation, Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Dongchuan Road 800, Shanghai 20040, China.

Email: jieyang@sjtu.edu.cn

Qi Yu, Department of Ophthalmology, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China.

Email: yu.qi@sjtu.edu.cn



## Introduction

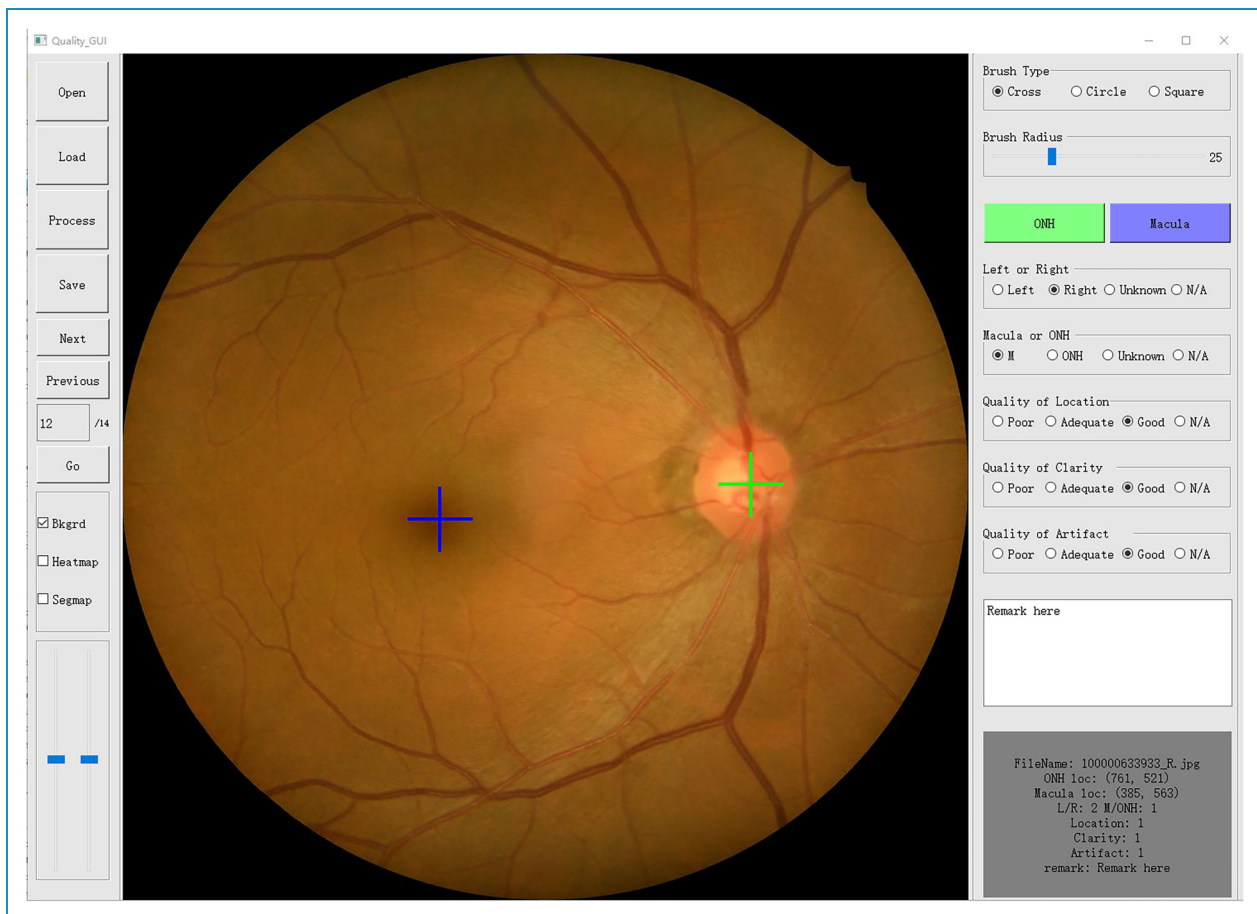
The number of people with visual-threatening eye diseases worldwide will be about 418 million by 2019 according to the World Health Organization (WHO) reports.<sup>1</sup> Screening and early interventions can reduce the risk of vision loss. Color fundus photography (CFP) is widely used for large-scale screening and treatment follow-ups since it is non-invasive and cost-effective. CFP images, as shown in Figure 1, are also widely used in clinical research. However, the image qualities of CFP often fluctuate due to technical or patient conditions. More than 25% of the retinal images cannot be diagnosed accurately due to poor image qualities according to the UK BioBank,<sup>2</sup> which means the quality problem is very common and will affect the subsequent diagnosis of ophthalmologists and computer algorithms.<sup>3</sup> Therefore, the pre-diagnosis assessment of CFP image quality is necessary.

Traditionally, ophthalmologists or technicians evaluate the CFP image quality manually. Due to the inefficiency, fallibility, and tediousness of humans, an automated computer-aided evaluation system is urgently needed. Indeed, previous researchers have designed methods for

DL-based pre-diagnosis,<sup>3,4</sup> they utilize several networks to learn several sub-tasks, respectively, and train them separately. However, existing works have their limitations. Firstly, the detailed clinical 3-grade quality criterion<sup>5</sup> is not adopted in these works. They only evaluate the quality using two classes (readable and unreadable), which lack details. Secondly, existing works only explore several sub-tasks and design a set of networks to fit, which decreases completeness and efficiency.

Based on the “Guidelines for image acquisition and grading of diabetic retinopathy screening in China,”<sup>5</sup> the image quality assessment process is divided into three categories focusing, respectively, on the image location, artifact, and clarity. There are three levels of grading (1: Good, 2: Adequate, 3: Poor) in each category. Finally, an overall score is generated according to the grading of each category. The photographer can make appropriate adjustments accordingly to avoid these problems.

In this study, we aimed to develop a deep learning-based multi-task framework for the pre-diagnosis of large-scale screening fundus photography. The grading of the CFP



**Figure 1.** The important landmark structures in color fundus photography (CFP) image (figure inspired by Chalakkal et al.<sup>6</sup>).

image quality using the Chinese three-grade standard is conducted. Some auxiliary tasks that can assist the quality grading as the detection of the optic nerve head (ONH) and macula, the classification of the laterality, and field of view (FoV) are also included. Different modules are combined in one multi-task framework to make full use of the features, which can be more accurate and efficient. We train our framework using local internal data sets and test on both internal and external data sets to show generalizability. Our framework can be used before human or computer-aided diagnosis to evaluate the reliability. And can also be used during photography to guide the photographer to acquire a high-quality image. The implementation is available at GitHub\*.

## Materials and methods

In this retrospective study, we trained, validated, and tested a deep-learning model for CFP image quality evaluation. The study was approved by the institutional review board of Shanghai General Hospital and conducted in accordance with the tenets of the Declaration of Helsinki. The requirement for Written consent was waived by the institutional review board because of the retrospective nature of the study. All analyzed data were anonymized and de-identified.

### Image acquisition

The data used in this work contain a total of 55,931 CFP images and were collected from three databases. They are the Shanghai Diabetic Retinopathy Screening Program (SHDRS) database, the Metabolic Management Center (MMC) database, the Eye-Quality (EyeQ)<sup>7</sup> database, the Diabetic Retinopathy Image Database (DRIMDB),<sup>8</sup> the Indian diabetic retinopathy image data set (IDRID)<sup>9</sup> database, and the MESSIDOR<sup>10</sup> database.

SHDRS database contains 21,180 retinal photographs taken by Topcon TRC-50DX with 45° FoV; the MMC database consists of 4049 retinal photographs taken by Topcon TRC-NW8 with 45° FoV; EyeQ<sup>7</sup> is a publicly available data set used for the retinal image quality assessment task. It is a subset of EyePACS data set<sup>11</sup> and is annotated by Fu et al.<sup>7</sup> It totally contains 28,792 retinal images. The images in this database are taken by multiple fundus cameras such as Topcon, Canon, Optovue, and so on. They are collected from multiple data centers. DRIMDB<sup>8</sup> is another publicly available data set that consists of 125 good-quality images and 69 poor-quality images. They were captured by a Canon CF-60UVi fundus camera. IDRID<sup>9</sup> contains 516 images which were acquired using a Kowa VX-10 alpha digital fundus camera with 50-degree FoV. MESSIDOR<sup>10</sup> consists of 1200 images which were acquired by Topcon TRCNW6 using 45°. Both IDRID and MESSIDOR are publicly available data sets commonly

used for grading diabetic retinopathy. All images in these two data sets are considered to be of high quality.<sup>12,13</sup>

The MMC and SHDRS data sets are used as the internal data set. Precisely, 20% of the internal data are used for internal validation. The EyeQ data set is used for external validation to evaluate the generalizability. The IDRID and MESSIDOR data sets are used externally to discuss the result of our model on good-quality images. The DRIMDB data set is used externally to evaluate the performance for transferring our criterion to the two-grade criterion.

### Ground truth labeling

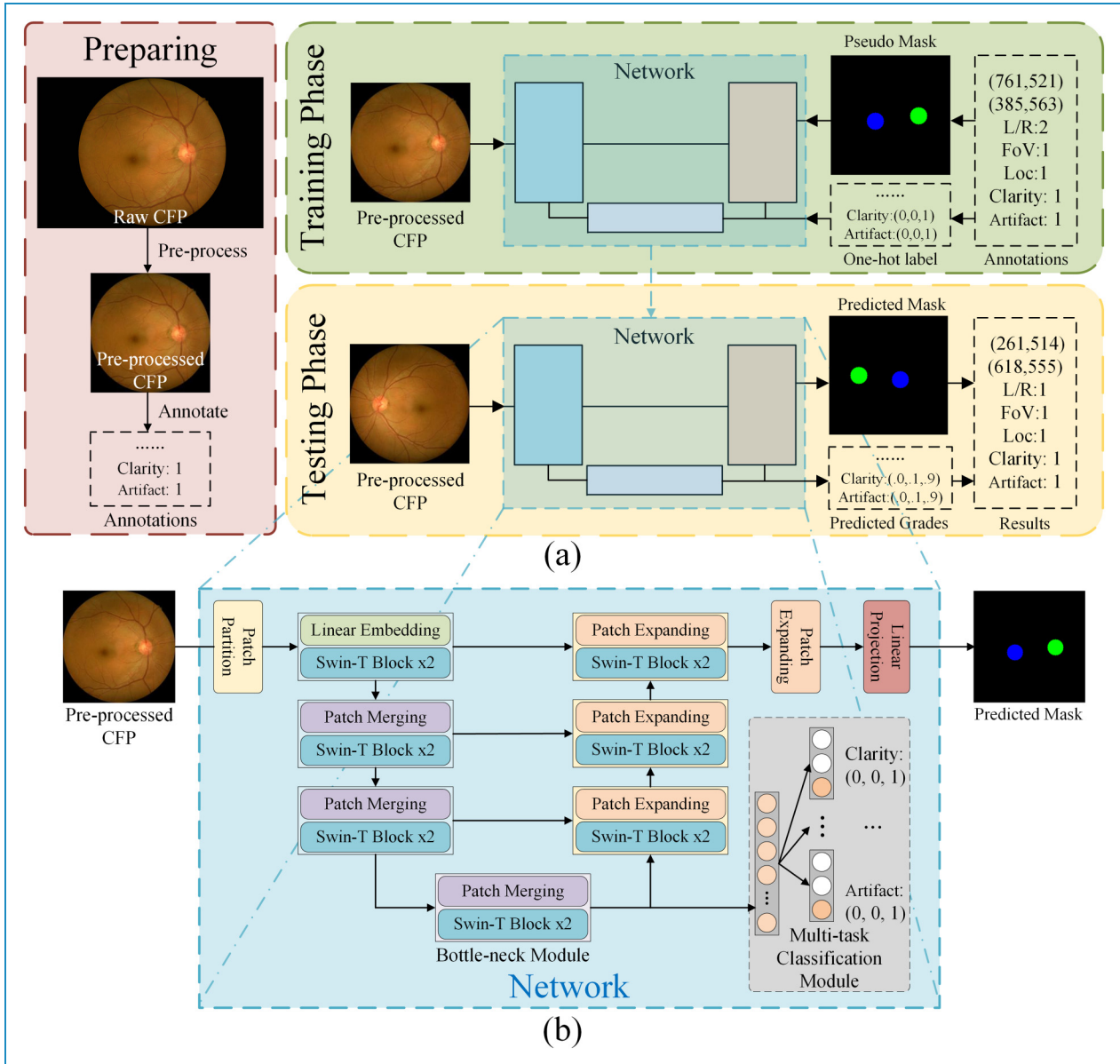
We are aiming at developing an automatic pre-diagnosis system with retinal image quality grading. Some auxiliary tasks that can assist the quality grading are also supposed to be included. Therefore, we annotate (1) the quality of the image and auxiliary attributes including, (2) the locations of ONH and macula, and (3) the FoV and laterality.

Eight licensed ophthalmologists (each with > three years of experience) and four retinal specialists (each with > 10 years of experience) were involved in the labeling process. Each fundus photograph was randomly assigned to two qualified ophthalmologists for annotation and then reviewed by a retinal specialist. Expert arbitration was done by the retinal expert if disagreement occurred in the previous process.

To assist ophthalmologists in annotating, we designed a software interface that includes all items in our pre-diagnosis system using PyQt5. Figure 2 shows the interface of the software. On the right side of the interface are the label tools, which are used to annotate the location of ONH and macula, the laterality, the FoV, and the quality of the image, respectively.

**Image quality labeling.** According to the “Guidelines for Image Acquisition and Reading of Diabetic Retinopathy Screening in China,”<sup>5</sup> the quality should be evaluated for three aspects: “Location,” “Clarity,” and “Artifact.” Each aspect should be evaluated using three-scale standards with detailed criteria (3: Poor, 2: Adequate, and 1: Good).

The “Location” sub-item is to assess if the FoV taken is standard. The key landmark (ONH or macula) should be located at the center of the image, that is, the distance between the key biomarker is expected to be less than the diameter of ONH (1PD). The clarity issue is usually caused by out-of-focus or movement during photoing. The anatomical structures as tiny vessels are very sensitive to clarity. Therefore, we evaluate the “Clarity” sub-item by observing the clarity of anatomical structures and details. During the photographing process of the non-mydratric fundus camera, due to the small pupil or the pupil contraction at the moment of exposure, the peripheral area of the fundus image may be covered by artifacts and the macular area may be dark, both of which are unreadable regions. The “Artifact” sub-item is to assess the readable-



**Figure 2.** The software interface we developed in this work. On the right side of the interface, the label tools, are used to annotate the location of ONH and macula, the laterality, the FoV, and the quality of the image, respectively. ONH: optic nerve head; FoV: field of view.

region percentage of one photograph. The overall quality is also three-grade and is evaluated by considering all these three sub-items.

The quantitative evaluation criteria are summarized in Table 1.

**Auxiliary attributes labeling.** Human experts annotate the locations of ONH and macula first and annotate the laterality and the FoV subsequently. Our software records these two locations by clicking the ONH and macula centers respectively. For the laterality and FoV, our software provides radio buttons to choose from. The “Left or Right” radio button group is to annotate the left and right

eyes. The “Macula or ONH” radio button group is to annotate the macula-centered FoV and optic-disc-centered FoV.

### Method procedures and network designing

**Procedures.** Our method procedures consist of the *Preparing Phase*, *Training Phase*, and *Testing Phase*, which are shown in Figure 3(a). In the *Preparing Phase*, the photographs collected from different data centers may have different resolutions and FoV masks. To reduce the influence of these variant factors and improve the efficiency of training, we pre-process the raw CFP image to be the same resolutions and FoV masks. The pre-processing of CFP images includes

**Table 1.** The details about the criteria of quality evaluation. PD denotes the diameter of the ONH.

Criteria for CFP image quality evaluation	
Location	Good (Grade 1): The distance between OD/macula center and image center < 1 PD
	Adequate (Grade 2): The distance between OD/macula center and image center > 1 PD and < 2PD
	Poor (Grade 3): The distance between OD/macula center and image center > 2 PD
Clarity	Good (Grade 1): Anatomical structures and details are clearly visible.
	Adequate (Grade 2): Anatomical structures and details are kind of blurred but acceptable.
	Poor (Grade 3): Anatomical structures and details are very blurred and unrecognizable.
Artifact	Good (Grade 1): The readable region covers 100% of the whole image
	Adequate (Grade 2): The readable region covers > 80% of the whole image
	Poor (Grade 3): The readable region covers < 80% of the whole image
Overall	Good (Grade 1): All the three sub-items are “Good”
	Adequate (Grade 2): Between “Poor” and “Good”
	Poor (Grade 3): At least one of the three sub-items is “Poor”

ONH: optic nerve head; CFP: color fundus photography; OD: optic disc; PD: pupillary distance.

cropping, padding, and resizing, which are detailed in the Supplemental Appendix. The pre-processed CFP image is then annotated by graders as mentioned above.

The *Training Phase* will start after acquiring the pre-processed CFP images and the corresponding annotations. In this phase, we generate the pseudo-ONH and macula segmentation masks by referring to their locations, which are detailed in the Supplemental Appendix. The laterality, the FoV, the quality of location, the clarity, and the readability annotations are transferred to one-hot encode. The processed CFP images with annotations are then fed into our deep network to train. We use label smoothing and cost-sensitive regularization strategies during training as in our previous work.<sup>14</sup>

The well-trained network will be evaluated in the *Testing Phase*. One unseen image is fed into our network, and then its ONH and macula segmentation results, the laterality, the FoV, the grades of “Location,” the grades of “Clarity,” and the grades of “Artifact” will be predicted. And finally, the other attributes including the location of ONH and macula are computed.

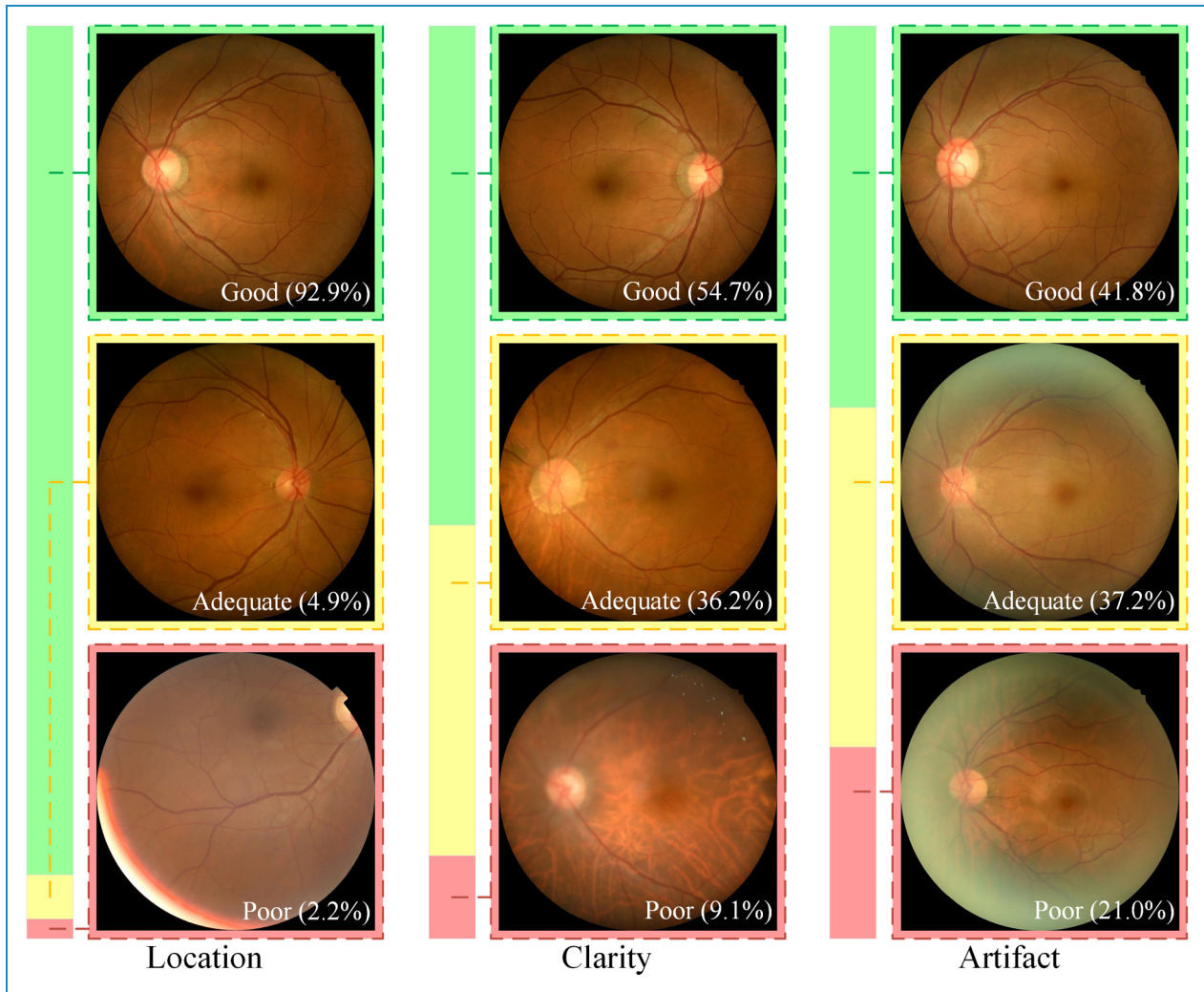
**Network designing.** Current researchers have paid attention to designing a set of deep neural networks to model these tasks.<sup>3,4,15–24</sup> However, aligning a set of different models in pre-diagnosis is neither elegant nor efficient, since models have to be run separately many times. Besides, the

potential relationships between these items are ignored. For example, the macula region is highly related to “clarity.”

To fully learn the image features, leverage the potential relationships between these items, and improve efficiency, we develop a multi-task deep neural network model to learn and predict these items simultaneously. We introduce the novel Swin-Unet<sup>25</sup> with transformer blocks as our backbone. The decoder module is to output the ONH and macula segmentation results. The multi-task classification module is designed and added to the bottleneck module of our backbone to output the grading results. The multi-task classification module first flattens the feature maps from the bottleneck of Swin-Unet, and then the flattened embedding connects to five parallel linear layers with three neurons. These five parallel linear layers are then activated by the softmax function to predict the one-hot grades of the laterality, the FoV, the grades of “Location,” the grades of “Clarity,” and the grades of “Artifact,” respectively. The network structure is shown in Figure 3(b).

### Statistical analysis

For the CFP image quality grading task, we evaluate the area under the receiver operating characteristic curve (AUROC), sensitivity, specificity, accuracy, and quadratic weighted kappa. The sub-items in the internal data set are calculated separately.



**Figure 3.** The whole procedure and the designed network are in this work. (a) The procedure. In the *Preparing Phase*, we pre-process raw color fundus photography (CFP) images and annotate them. The grading annotations and pseudo-segmentation masks are generated to train our network in the *Training Phase*. And the attributes are predicted and computed finally in the *Testing Phase*. (b) The network we designed. We use Swin-Unet as our backbone and then add a multi-task classification module to the bottle-neck module. Swin-Unet is to output the segmentation results while the multi-task classification module is to output the grading results and auxiliary attributes.

For the ONH and macula localization, the average Euclidean distance (ED),  $0.25R$  criterion,  $0.5R$  criterion, and  $1R$  criterion are evaluated, where  $R$  is the radius of ONH.<sup>24</sup> Besides, we also plot the success-rate-error-distance curve and calculate the area under the curve. The detailed definition can be found in the Supplemental Appendix.

For the laterality and FoV classification, The AUROC, sensitivity, specificity, and accuracy are utilized to evaluate the results of classification (grading).

## Results

### Data distribution

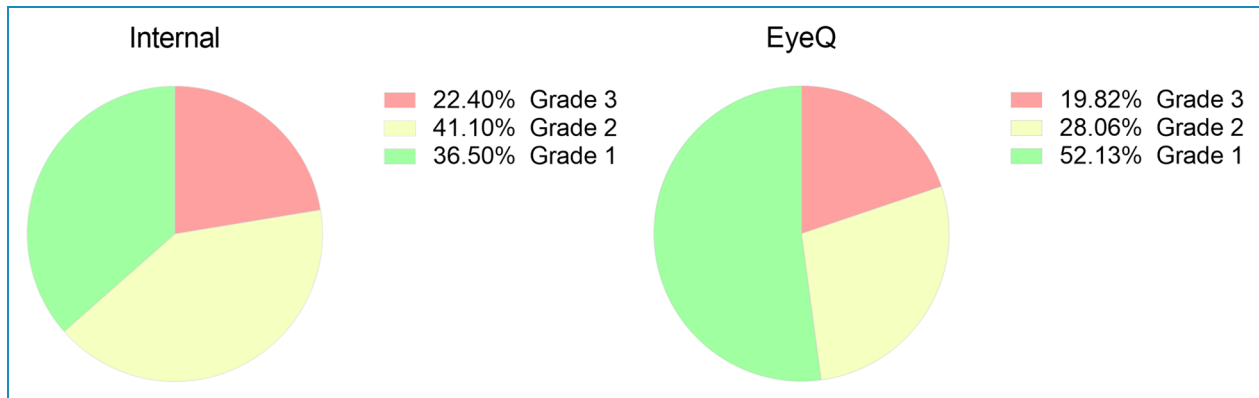
For the quality grading task, the data distribution of three quality subcategories is shown in Figure 4, where the

samples of each grade of each subcategory are given. The data distribution of the overall grade of the internal and external data sets is shown in Figure 5.

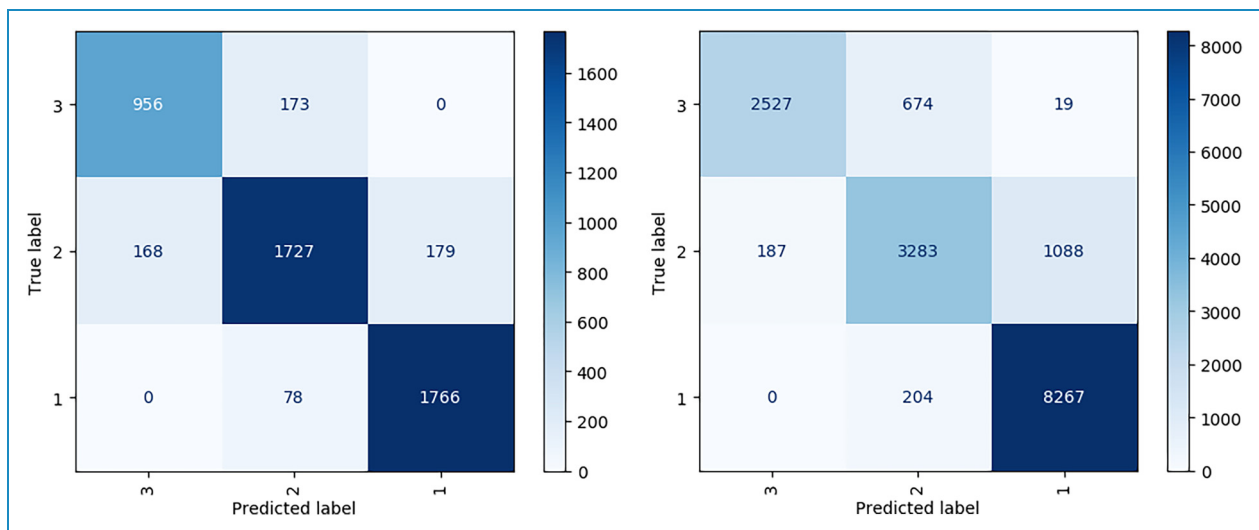
For the data in auxiliary tasks, the rate of left and right eye is about 1:1 for both internal and external sets. The rate of ONH FoV and macula FoV in these two data sets is all about 1:10.

### Quality grading

We evaluate the overall quality grades. The confusion matrix and the statistics are shown in Figure 6 and Table 2, respectively. On the internal set, the accuracy and Kappa scores of the three grades are 88.15% and 89.70%, respectively. The precision scores of the three



**Figure 4.** The data distribution of each quality subcategory with corresponding samples. One sample is given to each grade of each subcategory. The detailed percentage of each grade is shown in the bottom right-hand corner of the sample.



**Figure 5.** The data distribution of the overall grade. The left pie chart is our internal set while the right pie chart is the external Eye-Quality (EyeQ) set.

grades reach 85.05%, 87.31%, and 90.80%, with recall scores of 89.52%, 84.35%, and 86.31%, respectively.

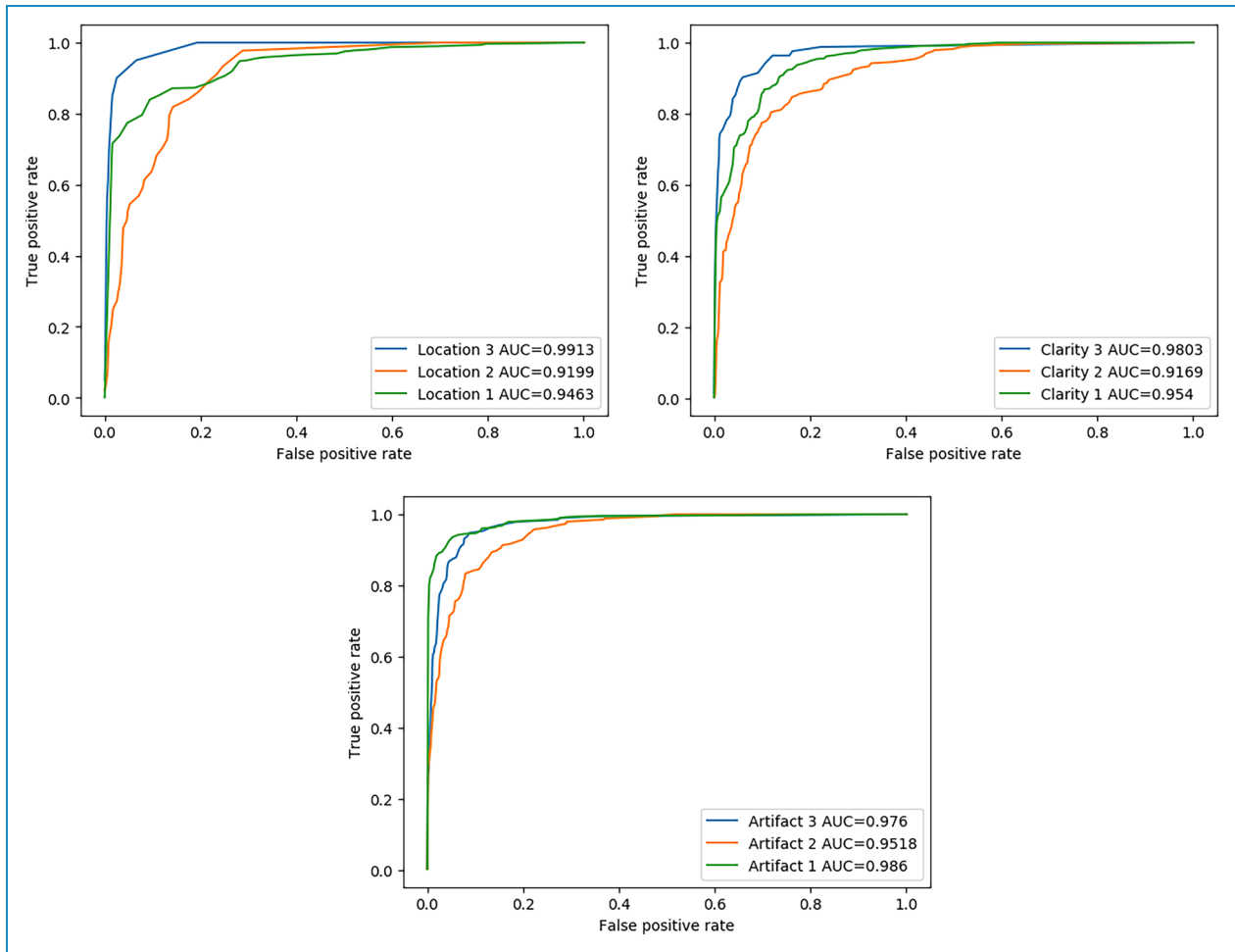
The overall quality grade is also evaluated on the external set. The confusion matrix and the statistics are shown in Figure 6 and Table 2, respectively. We achieved an accuracy of 86.63% and a Kappa of 88.55%. The precision scores of the three grades reach 93.11%, 78.90%, and 88.19% with recall scores of 78.48%, 72.03%, and 97.59%, respectively. Besides, the average precision and recall scores are also calculated using the mean value of three grades. The former reaches 86.76% and the latter reaches 82.70%.

The results of the quality grading of three subcategories on the internal data set are shown in Supplemental Figure 7 and Supplemental Table 3. We can see that for all the subcategories on the internal set, our model achieves AUROCs of between 0.920 and 0.990 mostly.

Especially, the accuracy of the “Location” sub-item on the internal set reaches 97.01% while the Kappa score is only 85.12%. This is because the distribution of grades is very imbalanced. The AUROCs of the three grades are 0.991, 0.920, and 0.946, respectively. The sensitivities reach 100.00%, 100.00%, and 87.13% with the specificities of 93.43%, 71.25%, and 89.06% on the three grades, respectively.

The “Clarity” subitem on the internal set achieves 85.27% and 82.28% accuracy and Kappa score, respectively. The AUROCs of the three grades are 0.980, 0.917, and 0.954, respectively. The sensitivities are 96.34%, 84.40%, and 90.89% with the specificities of 89.52%, 84.35%, and 86.31% on the three grades, respectively.

The “Artifact” sub-item on the internal set achieves 88.70% and 89.29% accuracy and Kappa score, respectively. The AUROCs of the three grades are 0.976, 0.952,



**Figure 6.** The confusion matrix of overall quality grading: (a) and (b) plot confusion matrix of internal and Eye-Quality (EyeQ) data sets, respectively.

**Table 2.** The results of quality grading on internal and external data sets. For two data sets, we report the precision and recall scores of each grade and the accuracy and Kappa scores of three grades.

Data set	Grade	Pre	Rec	Acc	Kappa
Internal	Grade 3	85.05%	84.68%	88.15%	89.70%
	Grade 2	87.31%	83.27%		
	Grade 1	90.80%	95.77%		
	Ave	87.72%	87.91%		
Eye-Quality (EyeQ)	Grade 3	93.11%	78.48%	86.63%	88.55%
	Grade 2	78.90%	72.03%		
	Grade 1	88.19%	97.59%		
	Ave	86.73%	82.70%		



and 0.986, respectively. The sensitivities are 95.26%, 89.88%, and 93.63% with the specificities of 91.30%, 86.60%, and 95.44% on the three grades, respectively.

### Auxiliary tasks

The localization results of ONH and macula and the classification results of FoV and laterality are shown in Tables 3 and 4, respectively.

Some key thresholds including  $0.25R$ ,  $0.5R$ , and  $1R$  are listed in Table 3, in which the average error distances in pixel and  $R$  is also reported. From Table 3, we can see that 98.12% of ONHs and 90.03% of the macula in the internal testing set are localized with distance error  $< 0.25R$ . More than 99% of ONHs and 97.5% of the macula in the internal testing set are localized with distance error  $< 1R$ . The average ONH and macula localization errors are about  $0.145R$  and  $0.313R$ , respectively. For the classification of FoV and laterality shown in Table 4, our model achieves AUROCs of 0.996 and 1.000 in the two tasks, respectively, in the internal validation. Both Tables 3 and 4 show that the performances of our model on the external EyeQ data set are close to that on

**Table 3.** The ONH and macula localization results of internal and external data sets.

	O/M	$R/4$	$R/2$	$1R$	AvgPix ( $\times R$ )
Internal	O	98.12%	99.00%	99.11%	10.840 (0.145R)
	M	90.03%	94.91%	97.56%	23.492 (0.313R)
EyeQ	O	97.19%	99.18%	99.51%	10.907 (0.145R)
	M	89.09%	96.51%	98.37%	19.796 (0.264R)

O and M denote the optic nerve head (ONH) and macula, respectively;  $R$  denotes the radius of ONH. The value in columns  $0.25R$ ,  $0.5R$ , and  $1R$  is the success rate, respectively. Column AvgPix( $\times R$ ) denotes the average error distance in pixel and  $R$  criterion. EyeQ: Eye-Quality.

**Table 4.** The results of FoV and laterality (abbreviated to “L”) classification on internal and external data sets. The sensitivity and specificity are chosen using the highest Yuden’s index. The accuracy is calculated with a threshold of 0.5.

	Attr.	AUROC	Sens.	Spec.	Acc.
Internal	FoV	0.996	97.87%	97.00%	98.10%
	L	1.000	98.88%	99.55%	99.00%
EyeQ	FoV	0.992	96.79%	95.07%	97.84%
	L	1.000	99.38%	99.68%	99.39%

EyeQ: Eye-Quality; FoV: field of view; AUROC: area under the receiver operating characteristic curve.

the internal set. More detailed figures can be found in Supplemental Figures S4 and S5.

### Discussion

In this study, we develop a module to assess the CFP image quality based on deep learning. The refined quality assessment criterion is introduced. The localization of ONH and macula and classification of FoV and laterality are included. The quality assessment of CFP images plays an important role in pre-diagnosis. Traditionally, ophthalmologists manually evaluate the CFP image quality first to assess the reliability of the diagnosis. The laterality and FoV may be evaluated additionally. Due to the inefficiency and subjectivity of manual evaluation, an automated system is needed. With the development of computer science, a number of researchers have explored the topic of retinal image quality assessment by using deep learning approaches, which are summarized in Supplemental Table 6. Some works related to the quality assessment as the classification of laterality and FoV are also summarized in Supplemental Table 6.

The related works mentioned in Supplemental Table 6 show promising performances. They mostly achieve accuracies of 80%–90% in the image quality assessment, and about 99% in the classification of FoV and laterality. Among them, Yuen et al.<sup>3</sup> concentrated on the completeness of the pre-diagnosis system. They develop a three-in-one pre-diagnosis system including the assessment of laterality, FoV, and quality. However, judging the CFP image quality only using “gradable and ungradable” is kind of coarse and subjective.

The CFP image quality evaluation process is divided into several categories according to the objective criteria. According to ‘Guidelines for image acquisition and reading of diabetic retinopathy screening in China’.<sup>5</sup> Three aspects affect the overall image quality including artifact, clarity, and image location. The artifact is usually caused by the small pupil or pupillary constriction. Low clarity is usually caused by optical path opacification or out-of-focus. The image location sub-item is to ensure the ONH (or macula) locates at the center of one photograph. Each of the categories should be evaluated using three grades to measure the extent of the problem. And the photographer can make solutions according to the type and the extent of issues. Therefore, introducing this quality evaluation criterion is of great clinical significance.

To address these issues, we developed a multi-task pre-diagnosis model. Our multi-task pre-diagnosis model shows several advantages. First, to the best of our knowledge, we are the first to introduce the three-grade detailed quality criterion to the deep learning task, which is more suitable for clinical usage. Second, our model contains all the processes of pre-diagnosis as localization of ONH and macula, the assessment of laterality, FoV, and quality,

which are all helpful to further diagnosis. Third, our model utilizes the multi-task strategy. All these indicators can be predicted simultaneously by a single forward pass, which increases efficiency. Besides, our multi-task strategy improves the performance of each task. The main reason is that multiple detailed information and supervisions are given, the network can have a comprehensive understanding of the retinal image, and the features can be made full use of. The results of individual training are shown in Supplemental Table S3.

Our pre-diagnosis model achieves promising performances. For the quality grading, our model shows mostly 0.92–0.98 AUROC for each grade. The overall accuracy of 88.15% and an overall Kappa of 89.70% are achieved on the internal data set, which shows high consistency. All the results on the external data set are very close to those on the internal data set, which shows the reliability and the generalization ability of our model. This also indicates that our algorithm can work on multiple types of fundus cameras. We also make a comparison with other deep-learning-based methods on the external EyeQ set as shown in Supplemental Table 7. For fair and comprehensive comparisons, we additionally calculate the precision, recall, and runtime here. Yu et al.<sup>18</sup> extract the feature embedding from saliency maps calculated by salient region detection, and the feature embeddings are then classified by SVM. Zago et al.<sup>17</sup> utilize the idea of fine-tuning or transfer learning in their quality assessment. We reproduce the methods of Yu et al.<sup>18</sup> and Zago et al.<sup>17</sup> on this data set. Fu et al.<sup>7</sup> designed a three-path convolutional neural network (CNN) to fuse the information from different color spaces. Muddamsetty and Moeslund<sup>26</sup> fuse manual and CNN features, and utilize random forest to classify quality. Their models are trained on the part of the EyeQ set while we only train our model by using local data. However, our model still shows competitive performance with accurate, precision, recall, and kappa scores of 86.63%, 86.73%, 82.70%, and 88.55%, respectively. As for the runtime, we also reproduce the model of Yuen et al.<sup>3</sup> We calculate the runtime of all the models under the same computing resource, and the results are also shown in Supplemental Table 7. The methods of Yu et al.<sup>18</sup> and Muddamsetty and Moeslund<sup>26</sup> partly need to be implemented on the CPU, which makes the runtime comparisons unfair. Our model only needs 4.554 ms for one prediction, which is the most efficient among these works. Compared with Yuen et al.,<sup>3</sup> which is the most similar to this work, our pre-diagnosis model is more efficient and detailed.

We also applied our three-grade quality model to the IDRID data set<sup>9</sup> and MESSIDOR<sup>10</sup> data set, which are generally considered to have no quality issues.<sup>12,13</sup> However, as indicated in Table 5, our model identified one image with poor “Location” quality and three images with poor “Clarity” on the IDRID data set. Furthermore, we observed

**Table 5.** The results of quality grading on the IDRID and MESSIDOR data sets. The total number of images and the count of images with corresponding quality issues are included.

Data sets	Grade	Location	Clarity	Artifact
IDRID (516)	Grade 3	1	3	0
	Grade 2	58	129	26
MESSIDOR (1200)	Grade 3	2	1	0
	Grade 2	44	106	48

DRIMDB: Diabetic Retinopathy Image Database; IDRID: Indian diabetic retinopathy image data set.

that 58, 129, and 26 images exhibited slight quality issues in the “Location,” “Clarity,” and “Artifact” sub-items, respectively. Similarly, on the MESSIDOR data set, two images have serious “Location” quality issues, and one image shows serious “Clarity” issues. We provide examples of such images in Figure 7. In the first column of Figure 7, images with “Location” issues are shown, where the macula centers are notably distant from the image centers. The second column shows images that are excessively blurred for accurate diagnosis, while the third column exhibits images with surrounding artifacts. Table 5 and Figure 7 indicate that our model is highly sensitive to quality issues.

Our three-grade quality model can be easily transferred to two-class (readable and unreadable) tasks. We calculate the maximum of the grade-3 dimension of the three sub-items. The results and comparison with others on the DRIMDB<sup>8</sup> database as shown in Table 6. Our model shows promising results, especially on the specificity, which means that our quality criterion is “stricter.”

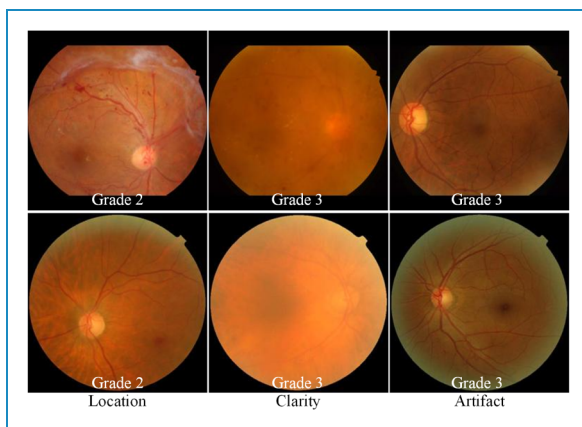
As for the auxiliary tasks. Our model achieves average error distances of 0.145R and 0.145R on internal and external data sets, respectively, during the localization of ONH. Our model achieves average error distances of 0.313R and 0.264R on these two data sets, respectively, during the localization of the macula. All the average localization error distances are less than 0.5R. For the classification of laterality and FoV, our model achieves over 0.99 AUROC scores on two data sets.

We also developed a user interface with PyQt5 to pre-diagnose intuitively. After loading the retinal images to be pre-diagnosed, the software can predict and present the ONH and macula localization results, laterality and FoV classification results, and quality grading results automatically. The other visualizing results including the ONH and macula segmentation map and heat maps can be shown by the user setting manually, which can give some references for the decision-making. The annotation functions are on the right side of the interface. As described in the

**Table 6.** The results comparison of quality grading on the DRIMDB data set. The sensitivity, specificity, and accuracy are chosen using the highest Yuden's index.

Authors	AUROC	Sens.	Spec.	Acc.
Zago et al.	0.998	95.65	98.55	97.10
Chalakkal et al.	-	97.6	97.8	97.7
Karlsson et al.	0.999	99.3	95.8	98.1
Ours	0.999	97.60	100.00	98.45

DRIMDB: Diabetic Retinopathy Image Database; AUROC: area under the receiver operating characteristic curve.



**Figure 7.** The image with quality issues on the IDRID (first row) and MESSIDOR (second row) data sets. The grades that our model predicts are listed.

DRIMDB: Diabetic Retinopathy Image Database; IDRID: Indian diabetic retinopathy image data set.

section, users can modify and annotate the annotations manually based on the automatic predictions. The pre-diagnosis model will perform better and better after some of the man-machine interactive processes.

Previous lectures have proved that low image quality often leads to unreliable diagnosis.<sup>27</sup> We found that low image quality also affects the performance of pre-diagnosis including the localization of ONH and macula, and the classification of laterality and FoV. For example, in our internal validation set, the average error distance of ONH in the subset with the overall quality of grade 1 is only 3.626 pixels, while the average error distance of ONH in the subset with the overall quality of grade 3 reaches 32.034 pixels. The AUROCs of laterality and FoV classification in the subset with the overall quality of grade 1 all reach 1.0. More details about the results under different quality conditions are shown in Supplemental Tables S1 and S2. The ratio of low-quality images on the internal data set is larger than that on the external data set. Therefore, some

of the indicators on the internal data set are lower than the external data set.

For the sub-items of quality. The location grade can be acquired by measuring the location of ONH and macula, which is completely objective and accurate. The criteria of artifact include the quantitative description, that is, the ratio of unreadable regions. The artifacts and dark regions that affect readability are usually conspicuous. Therefore, the readability grade is relatively objective and easy to learn. However, the criteria for clarity are completely qualitative. The label of clarity grade may be more subjective and inaccurate, which means the clarity grade is difficult to fit for the network. These are reflected in the results of quality grading on internal data sets (Supplemental Table 3), in which the clarity item shows the lowest accuracy and Kappa score.

The main strengths of this work, besides the promising performances, are as follows: First, our model contains quality grading, localization of ONH and macula, and classification of FoV and laterality, which is complete for the pre-diagnosis. Second, we explored the sub-items of quality grading, which is more objective and accurate. However, the limitation of this work exists. One limitation of this work is that our pre-diagnosis model may perform unstably when the image quality is extremely low, as shown in Supplemental Tables S1 and S2. Working robustly in low quality is still a problem that needs to be solved in the future. Besides, our work is mostly conducted on the 45°–50° FoV screening CFP images. The CFP images with other degrees of FoV should be evaluated in the future.

There are several potential applications of our pre-diagnosis image quality evaluation system. For photographers, real-time feedback can guide them to adjust the focus, exposure, and location to avoid technical issues. For ophthalmologists, our system can assist them to evaluate the image quality and reliability. For AI researchers, our system can be used in pre-classification or pre-processing before the downstream tasks. For instance, the localization of ONH can be used for further ONH segmentation, and the quality grading can filter out the images with good quality to reduce over-enhancement in quality enhancement tasks. Some recent works aimed at enhancing the image quality and proved that quality enhancement can boost the performance for downstream tasks such as disease grading and structure segmentation.<sup>28,29</sup> However, there is still a performance gap between the high-quality image and the enhanced image. Therefore, quality assessment and real-time feedback during photography are of great importance.

## Conclusion

In this article, we propose a refined image quality assessment module for color fundus photography using deep learning. The detailed criterion of CFP image quality

including three subcategories with three levels of grading improves the objectiveness and repeatability of the results. Auxiliary tasks including the localization of ONH and macula, the classification of laterality, and field of view are also contained in this module, which assists the quality assessment and improves the completeness of pre-diagnosis. The module was validated in both internal and external data sets with promising results, which shows the strength and generalizability of our module. This module can be used to assist photographers in getting better CFP images while imaging, filtering out the CFP images with poor image qualities that may affect further analysis, and pre-processing CFP images for downstream tasks.

**Acknowledgements:** The authors thank the study participants and their families for their approval and contribution to the research project.

**Availability of data and materials:** Data are available from the corresponding author upon reasonable request.

**Contributorship:** TG contributed the central ideas, analyzed most of the data, and wrote the initial draft of the paper. QY, KL, and HZ labeled the image and analyzed the data. QY designed the study and interpreted the data. XX and JY refined the ideas and revised the manuscript.

**Consent for publication:** Yes.

**Declaration of conflicting interests:** The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Ethics approval and consent to participate:** The study was approved by the institutional review board of Shanghai General Hospital and conducted in accordance with the tenets of the Declaration of Helsinki. The requirement for Written consent was waived by the institutional review board because of the retrospective nature of the study. All analyzed data were anonymized and de-identified.

**Funding:** The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research is partly supported by Shanghai Agriculture Applied Technology Development Program, China (grant no. 2019-02-08-00-08-F01118), Committee of Science and Technology, Shanghai, China (no. 19510711200), National Key R&D Program of China (2019YFB1311503), Shanghai Sailing Program (20YF1420800), and NSFC (61661010, 61977046, 62003208, 81600776, and 81970846).

**Guarantor:** TG is the guarantor for this article.

**ORCID iD:** Tianjiao Guo  <https://orcid.org/0000-0002-3186-0762>

**Supplemental material:** Supplemental material for this article is available online.

## References

1. Organization GWH. World report on vision, 2019. <https://www.who.int/publications/i/item/world-report-on-vision/>.
2. MacGillivray TJ, Cameron JR, Zhang Q, et al. Suitability of UK biobank retinal images for automatic analysis of morphometric properties of the vasculature. *PLoS One* 2015; 10: e0127914.
3. Yuen V, Ran A, Shi J, et al. Deep-learning-based pre-diagnosis assessment module for retinal photographs: a multicenter study. *Transl Vis Sci Technol* 2021; 10: 16–16.
4. Rim TH, Da Soh Z, Tham YC, et al. Deep learning for automated sorting of retinal photographs. *Ophthalmol Retina* 2020; 4: 793–800.
5. Fundus Disease Group of Chinese Ophthalmological Society EFDGobCMA E F D C. Guidelines for image acquisition and interpretation of diabetic retinopathy screening in China (2017). *Zhonghua Yan Ke Za Zhi* 2017; 53: 890–896.
6. Chalakkal RJ, Abdulla WH and Hong SC. Fundus retinal image analyses for screening and diagnosing diabetic retinopathy, macular edema, and glaucoma disorders. *Diabetes and fundus OCT* 2020; 1: 59–111. Elsevier.
7. Fu H, Wang B, Shen J, et al. Evaluation of retinal image quality assessment networks in different color-spaces. In: *International conference on medical image computing and computer assisted intervention*. Shenzhen, China: Springer, 2019, pp.48–56.
8. Sevik U, Köse C, Berber T, et al. Identification of suitable fundus images using automated quality assessment methods. *J Biomed Optic* 2014; 19: 046006–046006.
9. Porwal P, Pachade S, Kamble R, et al. Indian Diabetic retinopathy image dataset (IDRiD): a database for diabetic retinopathy screening research. *Data* 2018; 3: 25.
10. Decenci ere E, Zhang X, Cazuguel G, et al. Feedback on a publicly distributed image database: the Messidor database. *Image Anal Stereol* 2014; 33: 231–234.
11. Foundation CH. Diabetic retinopathy detection - identify signs of diabetic retinopathy in eye images, 2015. <https://www.kaggle.com/c/diabetic-retinopathy-detection/>.
12. Abdel-Hamid L. Retinal image quality assessment using transfer learning: spatial images vs. wavelet detail subbands. *Ain Shams Eng J* 2021; 12: 2799–2807.
13. He C. Application of deep learning and transfer learning in retinal image quality assessment. 2022.
14. Guo T, Liang Z, Gu Y, et al. Learning for retinal image quality assessment with label regularization. *Comput Methods Prog Biomed* 2023; 228: 107238.
15. Coyner AS, Swan R, Campbell JP, et al. Automated fundus image quality assessment in retinopathy of prematurity using deep convolutional neural networks. *Ophthalmology Retina* 2019; 3: 444–450.
16. Chalakkal RJ, Abdulla WH and Thulaseedharan SS. Quality and content analysis of fundus images using deep learning. *Comput Biol Med* 2019; 108: 317–331.
17. Zago GT, Andreao RV, Dorizzi B, et al. Retinal image quality ~ assessment using deep learning. *Comput Biol Med* 2018; 103: 64–70.

18. Yu F, Sun J, Li A, et al. Image quality classification for DR screening using deep learning. In: *2017 39th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. Jeju, Korea (South): IEEE, 2017, pp.664–667.
  19. Liu P, Gu Z, Liu F, et al. Large-scale left and right eye classification in retinal images. In: *Computational pathology and ophthalmic medical image analysis*. Granada, Spain: Springer, 2018, pp.261–268.
  20. Jang Y, Son J, Park KH, et al. Laterality classification of fundus images using interpretable deep neural network. *J Digit Imaging* 2018; 31: 923–928.
  21. Bellemo V, Yip MYT, Xie Y, et al. Artificial intelligence using deep learning in classifying side of the eyes and width of field for retinal fundus photographs. In: *Asian conference on computer vision*. Perth, Australia: Springer, 2019, pp.309–315.
  22. Al-Bander B, Al-Nuaimy W, Williams BM, et al. Multiscale sequential convolutional neural networks for simultaneous detection of fovea and optic disc. *Biomed Signal Process Control* 2018; 40: 91–101.
  23. Meng X, Xi X, Yang L, et al. Fast and effective optic disk localization based on convolutional neural network. *Neurocomputing* 2018; 312: 285–295.
  24. Guo T, Liang Z, Gu Y, et al. Deep multi-task framework for optic disc and fovea detection. *J Electron Imaging* 2021; 30: 1–18.
  25. Cao H, Wang Y, Chen J, et al. Swin-Unet: Unet-like pure transformer for medical image segmentation. In: *Computer vision–ECCV 2022 workshops*. Tel Aviv, Israel: Springer, 2022, pp.205–218.
  26. Muddamsetty SM and Moeslund TB. Multi-level quality assessment of retinal fundus images using deep convolution neural networks. In: *16th International joint conference on computer vision theory and applications (VISAPP-2021)*. Online: SCITEPRESS Digital Library, 2021, pp.661–668.
  27. Beede E, Baylor E, Hersch F, et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In: *Proceedings of the 2020 CHI conference on human factors in computing systems*. New York, USA: Association for Computing Machinery, 2020, pp.1–12.
  28. Shen Z, Fu H, Shen J, et al. Modeling and enhancing low quality retinal fundus images. *IEEE Trans Med Imaging* 2020; 40: 996–1006.
  29. Li H, Liu H, Hu Y, et al. An annotation-free restoration network for cataractous fundus images. *IEEE Trans Med Imaging* 2022; 41: 1699–1710.
-